NASA-STD-3001 Technical Brief

Human-in-the-Loop (HITL) Test Subject Sample Sizes

OCHMO-TB-018

Relevant Technical Requirements

NASA-STD-3001 Volume 2, Rev C

[V2 3102] Human Error Analysis
[V2 3101] Iterative Developmental Testing
[V2 4102] Functional Anthropometric Accommodation
[V2 5007] Cognitive Workload
[V2 10200] Physical Workload
[V2 8058] Glare Prevention
[V2 10001] Crew Interface Usability
[V2 10002] Design-Induced Error
[V2 10003] Crew Interface Operability
[V2 10004] Controllability & Maneuverability
[V2 10047] Visual Display Legibility

Executive Summary

Human performance varies based on a user's unique experiences. Test subjects must be representative of the full range of potential crewmembers in both physical and cognitive aspects. Sample sizes for integrated Human-in-the-Loop (HITL) testing should look to incorporate enough test subjects to provide confidence in the statistic while covering the range of critical anthropometric dimensions needed for the tasks. *NASA strongly recommends HITL tests verifying human performance parameters utilize 10 test subjects when the metric is sensitive to a user's unique experience (usability, workload, error, etc.). NASA requires that metrics that are less affected by a user's unique experience (such as legibility and glare) or utilize an even more homogeneous population (certified test pilots for handling qualities) utilize a minimum of 5 test subjects.* Verification sample sizes consider end user population, statistical parameters, published data on the likelihood of finding errors, lessons learned and past experience conducting HITL testing for spaceflight, and expert judgment with NASA community buy-in. There are no requirements on sample size for developmental tests and they can typically use fewer subjects. However, more mature systems should use a larger sample size during testing.

Application

Identify the End User Population & Sample Size

As of January 2022, NASA had 44 active astronauts. Adding management and international active astronauts, this number approaches 100. While 100 is a small population, it is still not realistic to test the entire current population. Additionally, in many ways *the astronaut population is a more homogeneous sample pool* than the population as a whole. Thus, the statistical significance of using a sample of 10 subjects is much more representative of the overall population since the existing population of astronauts is so small. **HITLs must be designed to use a reasonable sample of test subjects that represents the population of end users in terms of experience with the task, environment, and system, as well as physical and cognitive capabilities.**

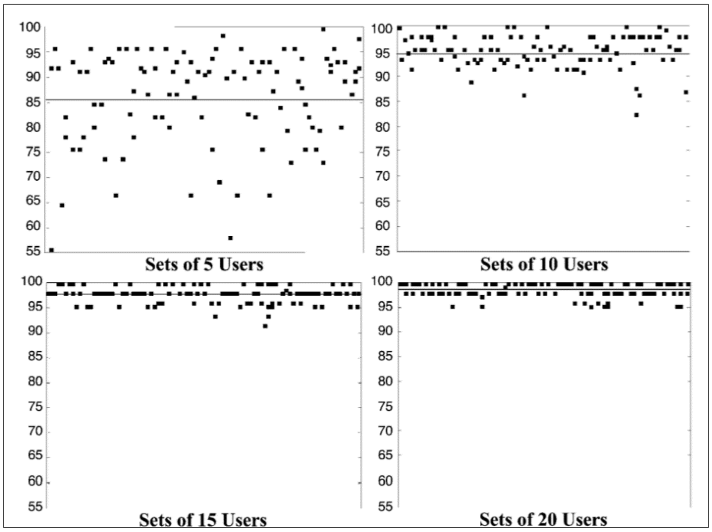
Increased Risk with Insufficient Sample

The area of highest concern is undiscovered design-induced errors. Sample size for errors encapsulates other human performance metrics as long as it also includes the range of critical anthropometric dimensions needed for the tasks.

# of Subjects	Likelihood of Error Detection	
	Easy to Find Errors	Hard to Find Errors
5	0.84	0.57
10	0.98	0.80
15	1.00	0.92
20	1.00	0.97

Nielsen and Landauer (1993)[5]

Increased Reliability of Results with Additional Users[1]



Application

The "Myth of 5"

There is only so low you can drive the subject sample size. While research has shown that the likelihood of detecting errors in a system increases with the number of test subjects, it is especially true for trying to identify less frequently occurring errors**. While five test subjects may be able to detect 84% of easy to find potential errors, those same five test subjects may only find 57% of the harder to find potential errors.[3]** Research also shows that testing more individuals increases the likelihood that you have detected harder to find errors.

Published data[1] for the likelihood of finding error in usability testing shows **that increasing the number of test subjects increases the likelihood of finding potential errors**. The gain from increasing from 5 to 10 test subjects is *noticeably greater* than the gain from 10 to 15 or 15 to 20.

Five is enough for <u>web</u> testing:

- when the original discount model for testing is followed
- when the results of testing are understood and clearly communicated
- when there is close cooperation between the client/sponsor and the test team
- **when the results are used for *diagnostic purposes* and *team learning***
- **when the expected result is insight, *not validation***

(Carol Barnum, Director of User Research and Founding Partner at UX Firm, LLC)[2]

"Five users can be enough sometimes. These are extremely rare. […] Where people have tested more than 5 users, it is absolutely clear that, unless problems are very easy to find, not only are more problems found, but also that the profile of problems as regards frequency and severity changes radically with further users. […] **Usability is about risk management. Risks diminish as we test more users**. The break even on cost-benefit is product specific. For some, one user is enough, but for others even 100 may be too few."

(Gilbert Cockton, North Umbria School of Design)[2]

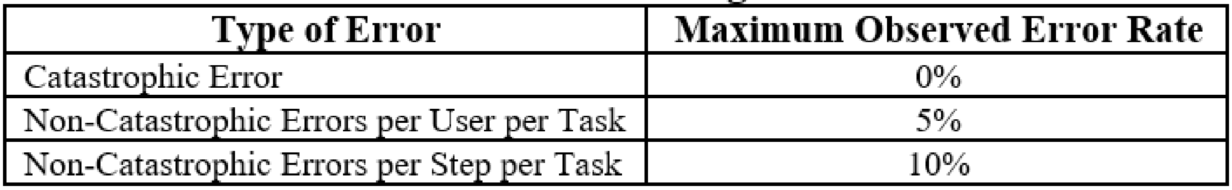
Types of Users

Proper testing needs test subjects with variability in familiarity with the design. Expert users have a bias, having already seen everything, so they are looking for a pattern. *Novice users expose things that an experienced user may not find.* Novice test subjects also expose errors that experienced users will make under stress. Testing with only expert users can *increase risk* that critical system errors have been overlooked. It is also inaccurate to assume that all users will have the same level of experience or training that the test subjects have – therefore a range of experience is recommended.[1]

Application

Implementation of **[V2 10002] Design-Induced Error:** The system shall provide crew interfaces that result in the maximum observed error rates listed in Table 29, Maximum Observed Design-Induced Error Rates.

Table 29—Maximum Observed Design-Induced Error Rates



Type of Error	Maximum Observed Error Rate
Catastrophic Error	0%
Non-Catastrophic Errors per User per Task	5%
Non-Catastrophic Errors per Step per Task	10%

For purposes of HITL testing, a scenario requiring evaluation will be defined as an activity driven by one or more related and sequential procedures. The procedure consists of a series of task steps, where a task step will be defined as a single instruction to the test subject, as is typical of current space flight procedures. Participants will maintain task completion times commensurate with the performance requirements.

- If any errors classified as having the potential of leading to a catastrophic outcome occur, the root cause of the error must be identified, mitigated satisfactorily (approved by NASA), and a re-test of the task performed to prove that the error has been eliminated.
- The percentage of errors (erroneous task steps) for each user shall be calculated by dividing the number of erroneous task steps and incomplete task steps by the total number of task steps and multiplying the result by 100.
- The percentage of users committing each error (erroneous task step) shall be calculated by dividing the number of users committing each erroneous task step by the total number of users and multiplying the result by 100].

Standard usability success criterion in industry is typically 80-95%; this requirement is more stringent. If more than 2 people in 20 have an issue on a step – it fails. The average number of procedural steps is 15; that would be less than one error per step per person.

When the standards and guidelines are implemented correctly, the usability engineer will:

- Review the verification plan.
- Perform adequate task analysis.
- Develop a usability test plan with accurately defined errors utilizing the appropriate number of test subjects.

Reference the OCHMO Usability, Workload, Error technical brief for additional information.

NASA recommends a sample size of 10 for easy and hard to find errors, with a mixed sample population of both experienced users and novice users to find the most potential errors possible. The majority of errors are design-induced and can be controlled for by improving design (i.e., human factors, crew interfaces). A small percentage of human errors are difficult to control for, including those caused by the user being distracted, fatigued, etc. This falls under the safety domain; good crew interface design cannot control for these types of errors.

Back-Up

Referenced Technical Requirements

NASA-STD-3001 Volume 2 Revision C

[V2 3006] Human-Centered Task Analysis Each human space flight program or project shall perform a human-centered task analysis to support systems and operations design.

[V2 3102] Human Error Analysis Each human space flight program or project shall perform a task-based human error analysis (HEA) to support systems and operations design.

[V2 4102] Functional Anthropometric Accommodation The system shall ensure the range of potential crewmembers can fit, reach, view, and operate the human systems interfaces by accommodating crewmembers with the anthropometric dimensions and ranges of motion as defined in data sets in Appendix F, Physical Characteristics and Capabilities, Sections F.2 and F.3.

[V2 5007] **Cognitive Workload** The system shall provide crew interfaces that result in Bedford Workload Scale ratings of 3 or less for nominal tasks and 6 or less for off-nominal tasks.

[V2 10200] Physical Workload The system shall provide crew interfaces that result in a Borg-CR10 rating of perceived exertion (RPE) of 4 (somewhat strong) or less.

[V2 8058] Glare Prevention Both direct and indirect glare that causes discomfort to humans or impairs their vision shall be prevented.

[V2 10001] Crew Interface Usability The system shall provide crew interfaces that result in a NASA-modified System Usability Scale (SUS) score of 85 or higher.

[V2 10002] Design-Induced Error The system shall provide crew interfaces that result in the maximum observed error rates listed in Table 29, Maximum Observed Design-Induced Error Rates.

[V2 10003] Crew Interface Operability The system shall provide interfaces that enable crewmembers to successfully perform tasks within the appropriate timeframe and degree of accuracy.

[V2 10004] Controllability and Maneuverability The spacecraft shall exhibit Level 1 handling qualities (Handling Qualities Rating (HQR) 1, 2 and 3), as defined by the Cooper-Harper Rating Scale, during manual control of the spacecraft's flight path and attitude when manual control is the primary control mode or automated control is non-operational.

[V2 10047] Visual Display Legibility Displays shall be legible in the viewing conditions expected during task performance

Reference List

1. Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers, 35*(3): 379-383.
2. Barnem, C., Bevan, N., Cockton, G., Nielsen, J., Spool, J., & Wixon, D. (2003, April 5-10). The "Magic Number 5": Is It Enough for Web Testing? [Conference Panel]. Computer Human Integration (CHI), Florida, USA. https://www.researchgate.net/publication/200553097_The_magic_number_5_is_it_enough_for_web_testing
3. Virzi, R.A. (1992). Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough? *Human Factors: The Journal of the Human Factors and Ergonomics Society, 34*(4): 457-468.
4. Nielsen, J., & Landauer, T.K. (1993). A Mathametical Model of the Finding of Usability Problems. *Conference on Human Factors in Computing Systems*. April 24-29, 1993. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.544.5899&rep=rep1&type=pdf